Source Controlled Operation of Non-Volatile Memories

Inventor: Raul-Adrian Cernea

FIELD OF INVENTION

[0001] The invention relates to programming floating gate memory cells, and in particular to improved techniques of reading and writing of dual floating gate memory cells.

BACKGROUND OF THE INVENTION

[0002] There are many commercially successful non-volatile memory products being used today, particularly in the form of small cards, which use an array of flash EEPROM cells. The basic architecture of an individual EEPROM memory cell includes source and drain diffusions, coupled by a channel, formed in a semiconducting substrate. The charge storage unit itself is positioned above the channel, separated from it by a dielectric layer. This charge storage unit is often referred to as a floating gate. Overlying the charge storage unit there can be a control gate, used to address the individual cell for programming and reading.

[0003] Some designs include a split channel architecture, as described, for example, in U.S. Pat. No. 5,095,344, granted to E. Harari, which is incorporated herein by this reference in its entirety. In a split channel cell both the charge storage unit and the control gate partially overlie the channel. This design has the advantage of simplicity, but the control gate performs the multiple functions of addressing the individual cells to be programmed or read, as well as the function of participating in the actual programming or reading of the cells.

[0004] Another design introduces an additional gate structure. This additional gate partially overlies the channel region and partially overlies the control gate. The portion that overlies the channel region forms a transistor that performs a selecting function. Therefore, this is generally called the select gate. The control gate is often called a “steering gate” in this arrangement. The select gate to carries out the addressing function and even may contribute to the programming, but the primary programming and reading functions are performed by the steering gate, as described, for example, in U.S. Pat. No. 5,313,421, granted to D. Guterman et al., which is incorporated herein by this reference in its entirety. This kind of memory cell is illustrated schematically in Fig. 1(a). The corresponding device structure is shown in Fig. 1(b).

[0005] Higher storage densities can be achieved by a dual cell design, as described, for example, in U.S. Pat. No. 5,712,180, granted to D. Guterman et al., which is incorporated herein by this reference in its entirety. There are two floating gates per cell in the dual cell architecture, overlying portions of the same channel. Corresponding steering gates are positioned above the floating gates. The select gate is formed above the steering gates and also overlies the channel itself. This type of memory cell is illustrated schematically in Fig. 2(a). The corresponding device structure is shown in Fig. 2(b).

[0006] In an array of cells the select gates of the memory cells along a row are usually coupled to form a wordline along that row. The diffusions in different rows are aligned and coupled to form bitlines that run along columns. Also, the steering gates in different rows are aligned and coupled to form steering lines that run along columns. A recent design of a memory cell array is described in U.S. Pat. No. 6,151,248, granted to E. Harari, which is incorporated herein by this reference in its entirety.

[0007] An alternative architecture is described in U.S. Pat. No. 6,091,633 granted to Cernea. Steering gates are connected along rows to form wordlines and select gates are connected together along columns to form bit selection lines. This is the reverse of the conventional arrangement. The diffusions in different rows are aligned and coupled to form bitlines that run along columns as in the conventional arrangement. This array architecture has certain advantages over conventional arrangements. The advantages of this arrangement as applied to embodiments of the present invention are further described in the description section.

[0008] Typically, each floating gate holds one bit of information. That is, the floating gate is either charged or not, representing a one or a zero. Alternatively, higher storage density may be achieved by using a range of charge levels to represent a range of memory states. Such a system is described in patent application number 09/793,370 (publication number 20020118574) by Gongwer, filed on February 26, 2001.

[0009] Flash EEPROM memories hold some key advantages over other types of memory systems. One of these advantages is the non-volatile nature of the data storage, making these systems prime candidates for a wide variety of applications, including digital cameras, recording music and utilization in mobile communications. Flash EEPROM is frequently used in memory cards that can be inserted or removed from such devices while maintaining the data stored in the memory.

[0010] A characteristic of flash memory systems is, however, the relatively long time it takes to program the cells. Programming can take as long as 10-1000 microseconds, longer than, for example, present day DRAMs.

[0011] Many cells are programmed simultaneously in the memory systems. The cells of the array are selected for programming in a certain scheme. The speed of the programming is influenced by this scheme. In some arrays the simultaneous programming is performed, for example, on every fourth, or every seventh cell only. Therefore these schemes require four or seven programming cycles to program all the cells of the array, respectively. One approach to increasing the speed of programming is to program adjacent cells at the same time. This is described, for example, in U.S. Pat. No. 6,493,269 by Cernea, which is incorporated herein by this reference in its entirety. However, the individual programming cycles may still be time consuming even if the number of cycles is reduced.

[0012] Thus, programming schemes that program memory arrays more rapidly than conventional schemes are desired.

[0013] Conventional techniques for programming memory cells use programming steps followed by verification steps to achieve the required memory state. Several such steps may be required to program a cell in this manner. This can be time consuming. Therefore, reducing the number of verification steps, or eliminating the need for verification, is desired.

SUMMARY OF INVENTION

[0014] A technique for programming memory cells incorporating charge storage units is disclosed. The technique involves generating hot electrons in a first transistor of the cell to charge the charge storage unit of a second transistor in such a way that the process is self-limiting. This means that when the charge in the charge storage unit reaches some predetermined level, charging stops or decreases to a very low level. The self-limiting effect is achieved by maintaining a constant current and voltage to the cell so that an increase in voltage across the second transistor as the charge storage unit charges results in a drop in voltage across the first transistor. The voltage across the first transistor produces hot electrons. Therefore, as the voltage across the first transistor drops, fewer hot electrons are produced. Eventually, the production of hot electrons stops or is reduced to a very low level. Thus, the process may self-limit at a charge level proportional to the voltage applied.

[0015] The final charge produced by the self-limiting process in the charge storage unit depends on the voltage across the cell during the charging process. By holding the voltage at one side of the cell at a fixed voltage, the final charge may be made dependant on the voltage at the other side. That is, a constant voltage is maintained at one side of the cell and a data-dependant voltage is supplied to the other side. The data-dependant voltage may represent a binary logic state or an analog logic state. The resulting charge in the charge storage unit may also represent either a binary or analog logic state.

[0016] Two and three transistor cells may be used according to different embodiments. The self-limiting technique may be used for a two transistor design as described above, or in a three transistor design by turning on the transistor that has a charge storage unit not being programmed so that it is not involved in the process.

[0017] A constant current may be provided by controlling the transistor that produces hot electrons. This is done by a current mirror circuit using a reference current. The resistance of this transistor is reduced by the current mirror to maintain a constant current as the resistance of the transistor with the charge storage unit increases as a result of the increasing charge in the charge storage unit.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1(a) illustrates schematically a two transistor memory cell of the prior art.

Figure 1(b) shows the structure of a two transistor memory cell of the prior art.

Figure 2(a) illustrates schematically a three transistor memory cell of the prior art.

Figure 2(b) shows the structure of a three transistor memory cell of the prior art.

Figure 3(a) illustrates an example of a memory system using the present invention.

Figure 3(b) illustrates schematically the configuration of a memory cell within an array of such cells.

Figure 4(a) shows a memory cell according to one embodiment of the present invention at the start of programming where hot electrons are generated and go to the charge storage unit.

Figure 4(b) shows a memory cell according to one embodiment of the present invention at the end of programming where hot electrons are no longer produced and the charge storage unit is charged.

Figure 5(a) illustrates schematically a two transistor memory cell according to an embodiment of the present invention with a current mirror circuit maintaining constant current and a programming voltage.

Figure 5(b) illustrates schematically a three transistor memory cell according to an embodiment of the present invention with a current mirror circuit maintaining constant current and a programming voltage.

Figure 6 illustrates a current mirror circuit according to an embodiment of the present invention.

Figure 7(a) shows memory cells according to an embodiment of the present invention at the start of programming.

Figure 7(b) shows memory cells according to an embodiment of the present invention at the end of programming.

Figure 7(c) shows memory cells according to an embodiment of the present invention during reading.

DETAILED DESCRIPTION OF EMBODIMENTS

[0018] An example of a memory system 300 incorporating the present invention is shown in Figure 3(a). This example shows an EEPROM cell array 310 with a bit line decoder 320 and a word line decoder 330. A bit select line decoder and control circuit 340 is connected to the array 310 and to the bit line decoder 320. A memory controller 350 is connected to the bit line decoder 320, word line decoder 330 and the bit select line decoder and control circuit 340. The memory controller 350 is also shown connected to a host by connecting lines 360. The host may be a personal computer, notebook computer, digital camera, audio player, various other hand held electronic devices, and the like. The memory system 300 of Figure 3(a) will commonly be implemented in a card according to one of several existing physical and electrical standards, such as one from the PCMCIA, the CompactFlash™ Association, the MMC™ Association, and others. When in a card format, the lines 360 terminate in a connector on the card that interfaces with a complementary connector of the host device. The electrical interface of many cards follows the ATA standard, wherein the memory system appears to the host as if it was a magnetic disk drive. Other memory card interface standards also exist. As an alternative to the card format, a memory system 300 of the type shown in Figure 3(a) may be permanently embedded in the host device.

[0019] Figure 3(b) shows an example of a cell 370 of the memory array 310. This example shows a three transistor cell. In the cell 370 of Figure 3(b) the bit select lines run parallel to the bit lines. This arrangement of bit lines and bit select lines is convenient for arrays using this invention.

[0020] Figures 4(a) and 4(b) illustrate an embodiment of the present invention where hot electrons from a first transistor 410 charge a charge storage unit 422 in a second transistor 420. The structure shown is part of a three transistor cell like that shown in Figure 3(b). It will be understood that the method shown is not limited to this particular structure but may be used in two transistor structures and other similar structures.

[0021] Figure 4(a) shows the situation at the start of the programming process. A data-dependant voltage V_S is applied to the source 413 of the first transistor 410. The drain 421 of the second transistor 420 is kept at a higher voltage V_D. This causes electrons to flow through the two transistors from left to right in Figure 4(a). The total current between the transistors 410, 420 is kept constant by a current mirror circuit (not shown in Fig. 4(a)) controlling the first transistor 410. Initially there is no charge in the charge storage unit 422. Therefore, the impedance of the second transistor 420 is low and the voltage is mainly across the first transistor 410. This is shown in the voltage profile of Fig. 4(a). The voltage across the two devices, $V_D - V_S$, is shown as being entirely across the first transistor 410 in this illustration. There is no voltage across the second transistor 420 because it is fully turned on as there is no charge on the charge storage unit 422 and the steering gate 423 above the charge storage unit 422 is held at a constant voltage during programming. The voltage on the steering gate 423 is selected to be close to V_D.

[0022] The high voltage gradient in the first transistor 410 produces hot electrons in the first channel region 412 under the first gate 411. Electrons are accelerated as they move through the first channel region 412. The acceleration they undergo is proportional to the voltage gradient. Electrons with sufficient energy to move from the first channel region 412 to the charge storage unit 422 are considered to be "hot electrons." Typically, this requires approximately 3.2 electron-volts of energy. This means that the electrons must pass through a voltage of about 3.2 volts in the first channel region 412 under the first gate 411 in order to have enough energy to reach the charge storage unit 422. If the voltage across the first transistor 410 is less than this then production of hot electrons is reduced. If the voltage across the first transistor 410 is increased above 3.2 volts the number of hot electrons increases. In the initial state of the cell, with no charge in the charge storage unit 422 and a sufficiently large value of $V_D - V_S$, hot electrons are produced in the first transistor 410 and are supplied to the charge storage unit 422 of the second transistor 420.

[0023] As electrons are supplied to the charge storage unit 422 it becomes negatively charged as shown in Fig. 4(b). This charge causes an increase in the impedance of the second channel region 424 under the charge storage unit 422. Because of this impedance, a voltage difference develops between the source 425 and drain 421 of the second transistor 420. This in turn causes a drop in the voltage between the source 413 and drain 414 of the first transistor 410. This is because the total voltage difference across the two transistors 410, 420 remains constant, $V_D - V_S$, and as the voltage across one transistor increases the voltage across the other drops. This is shown in the voltage profile of Figure 4(b) where V_1 is the voltage at a point between the two transistors 410, 420. Initially, the voltage at this point is equal to V_D because the first transistor 410 has a high impedance and the second transistor 420 has a low impedance. As the impedance of the second transistor 420 increases and the impedance of the first transistor 410 decreases, V_1 changes to some intermediate voltage between V_D and V_S. In the voltage profile of Fig 4(b) V_D - V_1 is the voltage across the second transistor 420 due to the charging of the charge storage unit 422. V_1 - V_S is the resulting voltage across the first transistor 410.

[0024] As the voltage difference between the source 413 and drain 414 of the first transistor 410 diminishes, the number of hot electrons produced also diminishes. This is because the reduced voltage gradient does not provide as many electrons with sufficient energy to reach the charge storage unit 422. Because the minimum energy needed to reach the charge storage unit 422 is approximately 3.2 electron volts, 3.2 volts is a cut-off voltage for hot electron production. When the voltage between the source 413 and drain 414 of the first transistor 410 reaches approximately 3.2 volts, hot electrons are no longer produced in the first transistor 410 and the charging of the charge storage unit 422 by hot electrons stops. Therefore, this process is self-limiting. Thus, the final charge does not depend on the length of time that the charging voltage is applied. The final charge is largely independent of this time. Therefore, precise timing is not needed. Also, since the charge is self-limited at the required level verification of the charge level is unnecessary, or at least, is not required as frequently as in previous techniques. It will be understood that the cut-off for production of hot electrons may not be abrupt and that charging of the floating gate may continue at a low level after this point is reached. However, the hot electron production is, at least, considerably reduced.

[0025] The charge stored in the charge storage unit 422 by the self-limiting process is a function of V_D - V_S. Here, V_D is held constant so the charge depends on V_S, the voltage at the source 413 of the first transistor 410. So, applying a particular voltage to the source 413 produces a predictable charge in the charge storage unit 422. The charge storage unit 422 continues charging until a sufficient charge builds up to stop the process. The charge needed to do this is a function of the voltage applied to the source V_S. The cut-off point for the self-limiting process comes when the voltage across the first transistor 410 reaches about 3.2 volts. At that point, the voltage across the second transistor 420 is V_D - V_S - 3.2. Therefore, the final charge on the charge storage unit 422 is proportional to V_D - V_S - 3.2. So, this charge stored can be set to different levels by varying V_S, the programming voltage used. Thus, the charging process is self-limiting at a charge level that corresponds to the value of V_S used. In a multi-state system the data-dependant voltage represents one of various possible logic states. In an analog system the data-dependant voltage may be any voltage from a continuum of possible voltage levels. Charge levels in the charge storage units may correspond to particular logic states, whether binary or analog.

[0026] The self-limiting aspect of this technique offers several advantages. Firstly, there is less need to verify the programmed charge level. The prior art typically uses a series of voltage pulses to charge the charge storage unit 422. Overcharging the charge storage unit 422 may occur if the voltage pulses are not stopped when the charge storage unit is at the required charge. To ensure that overcharging does not occur the charge in the charge storage unit 422 is checked during the programming routine in what is known as a verify step. This may be done several times during programming. This may be quite time consuming. The self-limiting technique may either eliminate the need for verification or reduce the need so that fewer verification steps are needed.

[0027] A second advantage is that several cells may be programmed in parallel to different levels using the self-limiting method. In the typical prior art technique, programming must be stopped at the critical time to prevent overcharging. If different cells are to be programmed to different levels their respective programming signals must be stopped at different times. This makes parallel programming difficult. In the present embodiment, each cell will stop charging when it reaches the appropriate level. Hence, a group of cells may be charged at the same time.

[0028] Memory cells may also be read in parallel with this type of cell. A constant voltage may be supplied to one side of the cells. The steering gate 423 corresponding to the charge storage unit 422 to be read is then held at a constant voltage. The other two transistors of the cell are turned on, that is, they are supplied with a sufficient voltage to produce a low impedance in the region below their respective gates.

[0029] Figure 5(a) shows one embodiment of the present invention. This embodiment has a two-transistor memory cell where one transistor has a charge storage unit 522. A current mirror circuit 530 is connected to the source 513 of the first transistor T1 to provide a constant current through the two transistors T1 and T2. The current mirror controls current by regulating the voltage to the gate 515 of T1 to vary the current flowing between the source 513 and drain 514 of T1. A programming state dependent voltage V_S is also connected to the source 513 of T1. This is used to program the charge storage unit 522 to the required state. A constant voltage V_D - V_S is maintained across the two transistors, T1 and T2. Also, the current between the transistors is constant. In other words, the total resistance of the two transistors is kept constant. The resistance of T2 is initially close to zero but increases as the charge storage unit 522 acquires charge. As T2 increases in resistance, the resistance of T1 is reduced by the current mirror circuit 530 as it maintains a constant current. As described above, this process is self-limiting because once the resistance of transistor T1 drops to a certain level, and the voltage between its source 513 and drain 514 drops accordingly, there comes a point where hot electrons are no longer produced and the charging of the charge storage unit 522 stops.

[0030] Figure 5(b) shows another embodiment of the present invention. This has a three transistor memory cell having two charge storage units 522. The current mirror circuit 530 provides a constant current through the three transistors. The current is controlled by controlling the voltage on the gate 515 of T1. In this embodiment only one transistor is programmed at a time. That is, when one of the transistors, T2 or T3 is being programmed the other is fully turned on so that it is fully conductive. For example, if T2 is to be programmed then T3 is turned on and becomes fully conductive. In this condition the circuit behaves like that of Fig. 5(a) with the current mirror 530 and V_S being connected to T1. A cross-section of the structure of a three transistor cell like that shown schematically in Fig. 5(b) is shown in Fig. 2(b).

[0031] Figure 6 shows a current mirror circuit 630 that may be used to provide constant current. The current mirror circuit 630 forms a current mirror when connected to the transistor T1. This current mirror has a reference current I_{ref} connected to two transistors, T1 and T4. One of the transistors, T1, is the transistor used to produce hot electrons. The other transistor T4 is chosen to be identical to T1. The current mirror circuit 630 ensures that the current flowing between the source 613 and drain 614 of T1 is the same as that flowing between the source 642 and drain 641 of T4. That is, the current is equal to the reference current I_{ref}.

[0032] Figures 7(a) and 7(b) show the programming and reading of cells in parallel. The cells may be connected as shown in Figure 3(b). Figure 7(a) shows the cells at the start of programming. On the top, V_D is applied to one side of each cell. This voltage is applied by the respective bitlines. For example, V_D may be 6.5 volts. The bit line at the other end of each cell is supplied with a programming voltage. In the example shown the programming voltages used are 0 volts, 1 volt and 2 volts. The transistor that is being programmed in each cell is the upper transistor 720 in Figure 7(a). The steering gate of transistor 720 is supplied with a constant voltage, VPGM. The steering gate of transistor 750 is supplied with an overdrive voltage. This is a voltage sufficient to turn on transistor 750 and so supply the programming voltage to the middle transistor 710. The gate of middle transistor 710 is controlled by the current mirror to provide a constant current through the cell, for example 1μA. At the start of programming, transistor 720 has no charge in its charge storage unit. Therefore, it has low impedance and the voltage at its source is approximately the same as the voltage at its drain, V_D.

[0033] Figure 7(b) shows the cells at the end of programming. The middle transistor 710 of each cell has a voltage of 3.2 volts between its source and its drain. Therefore, hot electrons are no longer produced and charging has stopped. Each of the programmed transistors 720 has a different voltage at its source and a different charge in its charge storage unit. Therefore, each transistor has a different threshold voltage.

[0034] Figure 7(c) shows the cells during reading. A constant voltage is applied to one side of the cells, V_D. This may be a different voltage to the voltage used for programming. For example, 3.2 volts may be used for V_D during reading. A constant voltage is supplied to the gate of the programmed transistor. This voltage is VPGM - 3.2 volts. Because the gate voltage is reduced by 3.2 volts from the gate voltage used during programming, the source voltage will also drop by 3.2 volts. This results in the voltage at the source of each programmed transistor being the same as the voltage used to program the cell.

[0035] Other embodiments of the present invention are possible involving a hybrid of the self-limiting technique with other techniques for programming. For example, the self-limiting technique may be used to quickly program the cell close to its targeted level. The final programming may then be completed by conventional techniques. Conventional techniques use voltage pulses to program the charge storage unit. To ensure the required level of charge is achieved, the voltage is verified as it approaches the target. Therefore, programming involves alternating between applying a pulsed voltage to program, then reading the voltage to see if it is close to the target. This is time consuming but may be done very accurately. A hybrid technique still provides some of the time saving of the self-limiting technique but may also have some advantages of conventional techniques such as greater programming accuracy.

[0036] The embodiments described above may be combined with conventional techniques of programming, erasing and reading to improve the speed of the memory. These techniques are not limited to the specific embodiments described above but may be applied in other similar structures. The structures described are not limited to use with the specific methods shown. They may be used in other applications not disclosed here.